# BBN/LIMSI:
# Progress for RT03

**EARS RT-03 Workshop**
**Boston, MA**
**19-20 May 2003**



## The Past Year

- **Solid progress with respect to our goals**
- **Intra-team and inter-team collaborations have made their mark in a very positive manner**
- **Much remains to be done**

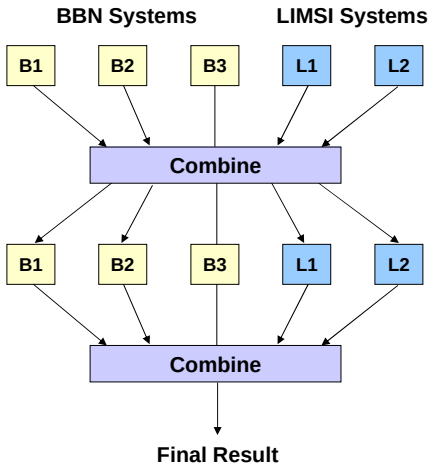
## Data Contributions

- CTS English: 80 hrs of CTRAN transcripts from Swbd2
- CTS Mandarin: CallFriend Dev set
- BN English Dev Set (LIMSI, BBN, SRI, CU)
- BN Mandarin Dev Set
- BN Arabic Dev Set





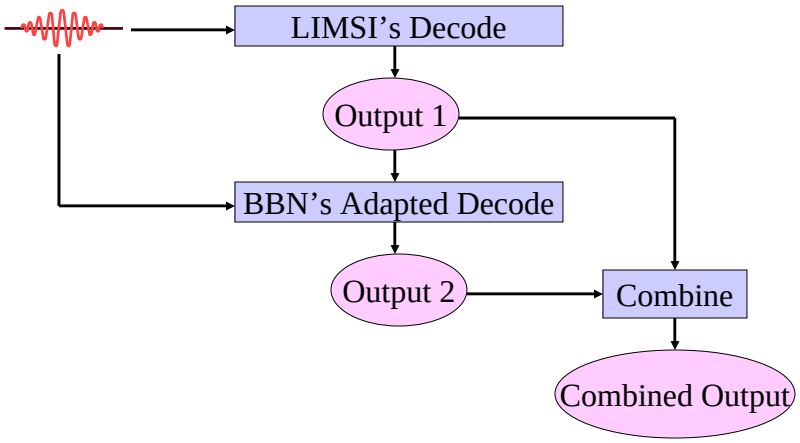
## CTS English Transatlantic System Architecture




Adapt to combined result and redecode

## BN English Post-Evaluation System Architecture

- **10xRT Condition**






BBN TECHNOLOGIES
A Verizon Company

## Outline

- **CTS Segmentation**
- **CTS LIMSI (English)**
- **Break**
- **CTS BBN (English, Mandarin, Arabic)**
- **BN BBN (English)**
- **BN LIMSI (English, Mandarin)**
- **BN BBN (Mandarin, Arabic)**



BBN TECHNOLOGIES
A Verizon Company

## Segmentation for Conversational Telephone Speech

Daben Liu, Francis Kubala



## Goal and Challenges

- Provide STT system automatic segmentation for conversational telephone speech (CTS)
  - For BBN and LIMSI
- Performance is measured by effect of segmentation on word error rate
- Two types of errors due to automatic segmentation
  - Insertion errors caused by noise and crosstalk
  - Deletion errors caused by missed speech

## Cross-Channel Event Modeling

- We model the following cross-channel events

| Events | Description |
|---|---|
| SS | Speech on both channels |
| SN | Speech on channel A and non-speech* on channel B |
| NS | Non-speech on channel A and speech on channel B |
| NN | Non-speech on both channels |

  * Non-speech includes silence, noise, crosstalk

- Advantages
  - Noise can be modeled explicitly
  - Crosstalk can be modeled by features from both channels



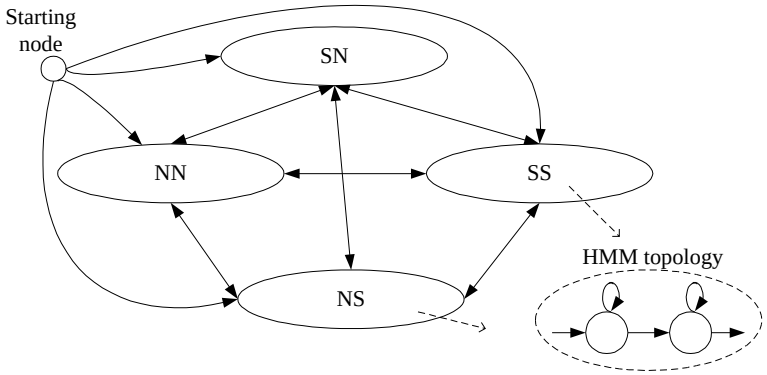


## Cross-Channel Event Modeling (Cont)

- Features
  - Concatenated features from both sides:
    14 MFCC and first order derivatives
  - Cross-channel features:
    - Energy difference (sigmoid function used to reduce dynamic range)
    - Cross-correlation coefficients (maximum delay 0.3 s)

- Training data
  - Balanced amount from SWB1, SWB2 cell, CallHome: total around 20 hours
  - Removed 800+ conversations with only one side transcribed
  - Removed NN segments > 2 s to avoid no-reference problem

## Ergodic Cross-Channel Event Network




- 512 component GMM for each HMM state
- Decode with an efficient Viterbi decoder
- Postprocess smoothing out any non-speech segment shorter than 0.1 s



BBN TECHNOLOGIES
A Verizon Company

## Results

**ML-based STT system with one-pass MLLR adaptation**

| WER | Manual segmentation (%) | Automatic segmentation (%) | Absolute difference (%) |
|---|---|---|---|
| English Eval 01 | 27.8 | 28.0 | +0.2 |
| English Eval 02 | 28.7 | 28.8 | +0.1 |
| Mandarin Eval97 | 46.0 | 46.3 | +0.3 |
| Arabic Eval97 | 51.1 | 51.5 | +0.4 |

**Final single MMI STT system**

| | | | |
|---|---|---|---|
| English Eval 02 | 24.0 | 24.4 | +0.4 |



BBN TECHNOLOGIES
A Verizon Company

## Conclusion


- We have developed an automatic segmentation for CTS based on cross-channel event modeling
- The WER from the automatic segmentation is about 0.1-0.4% higher than that from manual segmentation
- More work needs to be done to explain the widened gap for MMI system